Ilmenauer Beiträge zur Wirtschaftsinformatik

Herausgegeben von U. Bankhofer, V. Nissen
D. Stelzer und S. Straßburger

Udo Bankhofer, Dieter Joenssen

Hot-Deck-Verfahren zur Imputation fehlender Daten

Ergebnisse einer Simulationsstudie zur Untersuchung der Auswirkungen einer wiederholten Verwendung des Spenderobjekts

Arbeitsbericht Nr. 2011-05, Dezember 2011

Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften
Institut für Wirtschaftsinformatik

Autor: Udo Bankhofer, Dieter Joenssen

Titel: Hot-Deck-Verfahren zur Imputation fehlender Daten – Ergebnisse einer Simulationsstudie zur Untersuchung der Auswirkungen einer wiederholten Verwendung des Spenderobjekts

Ilmenauer Beiträge zur Wirtschaftsinformatik Nr. 2011-05, Technische Universität Ilmenau, 2011

ISSN 1861-9223

ISBN 978-3-938940-37-2

urn:nbn:de:gbv:ilm1-2011200564

Anschrift: Technische Universität Ilmenau, Fakultät für Wirtschaftswissenschaften, Institut für Wirtschaftsinformatik, PF 100565, D-98684 Ilmenau.
http://www.tu-ilmenau.de/wid/forschung/ilmenauer-beitraege-zur-wirtschaftsinformatik/

Gliederung

1 Problemstellung ... 1

2 Hot-Deck-Verfahren im Überblick ... 2

2.1 Sequentielle Hot-Deck-Verfahren ... 3

2.2 Simultane Hot-Deck-Verfahren ... 4

3 Stand der Forschung ... 4

4 Design der Simulationsstudie ... 5

4.1 Forschungsfragen ... 5

4.2 Einflussfaktoren ... 6

4.3 Gütekriterien ... 9

4.4 Durchführung der Studie ... 10

5 Ergebnisse der Simulationsstudie ... 11

5.1 Auswirkungen einer beschränkten Verwendung von Spenderobjekten ... 11

5.2 Analyse der Einflüsse auf die Spenderverwendungshäufigkeit ... 13

5.2.1 Analyse der Haupteffekte ... 13

5.2.2 Analyse von Wechselwirkungen ... 14

5.3 Analyse der Häufigkeit einer Verwendung der Spenderobjekte ... 16

6 Zusammenfassung und Ausblick ... 17

Literaturverzeichnis ... 19

Zusammenfassung: Hot-Deck-Verfahren sind spezielle, auf Imputationsklassen basierende Imputationsverfahren. Sie zeichnen sich dadurch aus, dass fehlende Daten durch vorhandene Werte der vorliegenden Datenmatrix imputiert werden. Das Objekt, das dabei die vorhandenen Daten zur Imputation liefert, wird als Spenderobjekt bezeichnet. Da die Ersetzung der fehlenden Daten eines Objekts durch die Ausprägungen eines ähnlichen Spenderobjekts sinnvoller ist die Ersetzung durch die Ausprägungen eines beliebig ausgewählten Spenderobjekts, erfolgt dieser Verdopplungsprozess innerhalb der zuvor gebildeten Imputationsklassen. Durch diese grundsätzliche Vordopplungseigenschaft der Hot-Deck-Verfahren stellt sich damit das Problem, dass ein Spenderobjekt wiederholt zur Imputation ausgewählt werden kann. Im Extremfall könnten somit alle fehlenden Werte bei einem Merkmal durch ein und dieselbe Ausprägung eines einzigen Spenderobjekts ersetzt werden. Aus diesem Grund erfolgt bei einigen Varianten der Hot-Deck-Verfahren eine Begrenzung der Anzahl, wie häufig ein Objekt als Spenderobjekt verwendet werden darf. Damit stellt sich zwangsläufig die Frage, unter welchen Bedingungen eine derartige Begrenzung überhaupt sinnvoll ist. Im Rahmen dieser Arbeit wird daher eine umfangreiche Simulationsstudie zur Beantwortung dieser Frage durchgeführt. Dabei zeigt sich, dass es deutliche Unterschiede zwischen Hot-Deck-Imputationen gibt, bei denen die Spenderverwendungshäufigkeit variiert wird. Darüber hinaus können auch einige Einflussfaktoren identifiziert werden, die für oder gegen eine Begrenzung der Spenderobjekte sprechen.

Schlüsselworte: Hot-Deck-Verfahren, fehlende Daten, Imputation, Simulationsstudie

1 Problemstellung

Fehlende Daten stellen ein häufig anzutreffendes Problem bei realen empirischen Untersuchungen dar. Im Fall fehlender Werte können die herkömmlichen, auf vollständigen Daten basierenden Auswertungsmethoden nicht mehr unmittelbar zur Anwendung kommen. Somit ergibt sich die Notwendigkeit einer expliziten Berücksichtigung der fehlenden Werte im Rahmen der Untersuchung.

Zur Behandlung fehlender Daten lassen sich im Wesentlichen drei grundlegende Strategien unterscheiden: Eliminierung, Imputation und direkte Auswertung der vorhandenen Daten. Während bei der direkten Auswertung der vorhandenen Daten die gewünschten Analyseergebnisse, wie beispielsweise Verteilungsparameter, unmittelbar auf Basis des unvollständigen Datenmaterials bestimmt werden, wird durch die Anwendung von Eliminierungs- und Imputationsstrategien eine vollständige und damit mit herkömmlichen Methoden auswertbare Datengrundlage bereitgestellt. Dabei werden im Fall einer Eliminierung Objekte bzw. Merkmale mit fehlenden Werten aus der Analyse ausgeschlossen und im Fall einer Imputation die fehlenden Daten durch geeignete Schätzwerte ersetzt (vgl. z.B. Kim, Curry, 1977, Allison, 2001).

Die vorliegende Arbeit setzt im Bereich der Imputationsverfahren an. Im einfachsten Fall können fehlende Daten durch einen geeigneten Lageparameter oder einen Verhältnisschätzer ersetzt werden. Darüber hinaus können auch multivariate Verfahren, wie beispielsweise die Regressions-, Varianz- oder Diskriminanzanalyse zur Anwendung kommen. Bei diesen Ansätzen werden verstärkt vorhandene Informationen zur Schätzung der fehlenden Daten herangezogen. Eine weitere Kategorie von Imputationsverfahren stellen Ansätze dar, die von sogenannten Imputationsklassen, d.h. Gruppen möglichst ähnlicher Objekte ausgehen. Dabei wird die durch die Klassifikation der Objekte vorliegende Information über die Ähnlichkeiten der Objekte zur Bestimmung von Imputationswerten herangezogen. Der Vorteil dieser Verfahrenskategorie liegt darin, dass weniger restriktive Annahmen über den Ausfallmechanismus vorliegen müssen. Während die anderen Imputationsverfahren im Allgemeinen die Eigenschaft MCAR (missing completely at random) bzw. unter gewissen Voraussetzungen MAR (missing at random) zwingend voraussetzen,[1] kann eine Imputation

[1] Eine ausführliche Darstellung dieser Ausfallmechanismen kann beispielsweise der Arbeit von Bankhofer (1995, S. 12-21) entnommen werden.

auf Basis von Imputationsklassen gegebenenfalls auch im Fall des Vorliegens der Eigenschaft NMAR (not missing at random) erfolgen (vgl. Bankhofer, 1995, S. 112-119, Andridge, Little, 2010, S. 49-51).

Bei den ausschließlich auf Imputationsklassen basieren Verfahren wird zwischen Cold-Deck- und Hot-Deck-Verfahren unterschieden. Während bei den Cold-Deck-Verfahren eine Imputation der fehlenden Daten durch Werte einer externen Quelle (z.B. ähnliche Untersuchungen oder zusätzliche Erhebungen) erfolgt, zeichnen sich die Hot-Deck-Verfahren dadurch aus, dass fehlende Daten durch vorhandene Werte der vorliegenden Datenmatrix imputiert werden. Das Objekt, das dabei die vorhandenen Daten zur Imputation liefert, wird als Spenderobjekt bezeichnet. Da die Ersetzung der fehlenden Daten eines Objekts durch die Ausprägungen eines ähnlichen Spenderobjekts sinnvoller ist als die Ersetzung durch die Ausprägungen eines beliebig ausgewählten Spenderobjekts, erfolgt dieser Verdopplungsprozess innerhalb der zuvor gebildeten Imputationsklassen.

Durch diese grundsätzliche Vordopplungseigenschaft der Hot-Deck-Verfahren stellt sich damit das Problem, dass ein Spenderobjekt wiederholt zur Imputation ausgewählt wird. Im Extremfall könnten somit alle fehlenden Werte bei einem Merkmal durch ein und dieselbe Ausprägung eines einzigen Spenderobjekts ersetzt werden. Aus diesem Grund erfolgt bei einigen Varianten der Hot-Deck-Verfahren eine Begrenzung der Anzahl, wie häufig ein Objekt als Spenderobjekt verwendet werden darf. Damit stellt sich zwangsläufig die Frage, unter welchen Bedingungen eine derartige Begrenzung überhaupt sinnvoll ist und ob gegebenenfalls ein jeweils geeigneter Wert hierfür bestimmt werden kann. Diesen Fragestellungen soll im Rahmen dieser Arbeit nachgegangen werden. Dazu werden in Kapitel 2 zunächst die grundlegenden Varianten der Hot-Deck-Verfahren im Überblick dargestellt. Das Kapitel 3 widmet sich anschließend dem aktuellen Stand der theoretischen und empirischen Forschung zu dieser Thematik. In Kapitel 4 wird dann auf das Design der Simulationsstudie ausführlich eingegangen und in Kapitel 5 erfolgt schließlich eine Darstellung der Ergebnisse der Studie.

2 Hot-Deck-Verfahren im Überblick

Grundsätzlich läuft eine Hot-Deck-Imputation in den folgenden Schritten ab: Zunächst werden die Imputationsklassen festgelegt. Anschließend wird innerhalb der Klassen für jede fehlende Merkmalsausprägung eine vorhandene Ausprägung bezüglich desselben Merk-

mals ausgewählt, so dass die fehlenden Daten schließlich durch die jeweils ausgewählten Werte imputiert werden können. Durch die unterschiedlichen Möglichkeiten der Auswahl von Imputationswerten aus den vorhandenen Daten sind auch unterschiedliche Varianten von Hot-Deck-Verfahren denkbar. In der Literatur (vgl. z.B. Kalton, Kasprzyk, 1982, Kalton, Kasprzyk, 1986, Brick, Kalton, 1996, Marker et al., 2002) wird dabei vor allem zwischen sequentiellen und simultanen Hot-Deck-Verfahren unterschieden.

2.1 Sequentielle Hot-Deck-Verfahren

Ausgehend von einer unvollständigen Datenmatrix werden bei den sequentiellen Hot-Deck-Verfahren in einem ersten Schritt innerhalb der festgelegten Imputationsklassen die Objekte in eine Reihenfolge gebracht und Startwerte für die zur Imputation heranzuziehenden Merkmalsausprägungen bestimmt. Abhängig von den verschiedenen Möglichkeiten, die Objektreihenfolge sowie die Startwerte festzulegen, ergeben sich unterschiedliche Varianten der sequentiellen Hot-Deck-Verfahren. Die Reihung der Objekte innerhalb der Imputationsklassen kann dabei zufällig, gemäß der Objektnummer oder nach der Ähnlichkeit der Objekte erfolgen (vgl. z.B. Little, Rubin, 1987, S. 65). Für die Festlegung der Startwerte gibt es in der Literatur (vgl. z.B. Little, Rubin, 1987, S. 65, Schnell, 1986, S. 109) ebenfalls mehrere Vorschläge. Danach können die Startwerte innerhalb der Klassen durch zufällige Auswahl einer vorhandenen Ausprägung für jedes Merkmal, durch die Klassenmittelwerte sowie durch die Anwendung eines Cold-Deck-Verfahrens ermittelt werden. Dabei erscheint die Verwendung der Klassenmittelwerte problematisch, da dadurch die Verdopplungseigenschaft der Hot-Deck-Verfahren verloren geht und somit keine klare Abgrenzung zur Mittelwertersetzung möglich ist.

Sind schließlich die Reihenfolge der Objekte und die Startwerte festgelegt, dann erfolgt eine sequentielle Abarbeitung der gesamten Datenmatrix in der Art, dass innerhalb der Imputationsklassen geprüft wird, ob die Objekte der Reihe nach eine fehlende Ausprägung bezüglich der einzelnen Merkmale besitzen. Ist dies nicht der Fall, dann wird die entsprechende Ausprägung des Objekts zum neuen Imputationswert für das gerade betrachtete Merkmal innerhalb der Imputationsklasse, andernfalls wird dem Objekt der für das betrachtete Merkmal vorliegende, aktuelle Imputationswert der Klasse zugewiesen.

2.2 Simultane Hot-Deck-Verfahren

Während bei den sequentiellen Hot-Deck-Verfahren jeweils merkmalsweise eine fehlende Ausprägung durch eine vorhandene ersetzt wird, erfolgt bei den simultanen Hot-Deck-Verfahren die Ersetzung sämtlicher fehlender Ausprägungen eines Objekts durch die Ausprägungen eines einzigen anderen Objekts.

Innerhalb der Imputationsklassen kann das Objekt, das die Imputationswerte liefert, entweder zufällig oder bewusst im Sinne einer größtmöglichen Ähnlichkeit zum unvollständig vorliegenden Objekt bestimmt werden. Durch die Verwendung unterschiedlicher Ähnlichkeitsmaße sind auch unterschiedliche Varianten dieses Verfahrens gekennzeichnet. In jedem Fall muss aber der Objektvektor, der die Imputationswerte enthält, vorhandene Daten bezüglich aller Merkmale aufweisen, die im Rahmen der Ersetzung relevant sind. Um dies zu gewährleisten und darüber hinaus eine einfache Anwendung der Verfahren zu ermöglichen, wird in der Literatur meist die Verwendung der vollständig vorliegenden Objekte zur Festlegung der Imputationswerte vorgeschlagen (vgl. z.B. Schnell, 1986, S. 110).

3 Stand der Forschung

Die Begrenzung der Häufigkeit, mit der ein Spenderobjekt im Rahmen einer Hot-Deck Imputation verwendet wird, ist zum ersten Mal Untersuchungsgegenstand in der Arbeit von Kalton und Kish (1981). Sie kommen aufgrund kombinatorischer Überlegungen zu dem Ergebnis, dass es bei den sequentiellen zufälligen Hot-Deck-Verfahren durch ein Ziehen des Spenderobjekts ohne Zurücklegen zu einer Reduktion der Varianz von Verteilungsparametern der vervollständigten Daten und damit zu einer Verbesserung der entsprechenden Schätzgenauigkeit kommt. Für eine Begrenzung der Verwendungshäufigkeit eines Spenderobjekts sprechen auch zwei weitere Gründe. Zum einen wird dadurch das Risiko begrenzt, im Extremfall ausschließlich ein einzigen Spender zu verwenden (vgl. Sande, 1983, S. 345). Zum anderen wird aber auch die Wahrscheinlichkeit reduziert, ein Spenderobjekt mit extremen Werten zu häufig zu verwenden (vgl. Bankhofer, 1995, S. 125, Strike et al., 2001, S. 893). Demgegenüber argumentieren Andridge und Little (2010, S. 43), dass eine Einschränkung der Spenderverwendungshäufigkeit zwangsläufig auch die Auswahl des jeweils ähnlichsten Spenderobjekts einschränkt und durch das Zulassen einer häufigeren Verwendung eines Spenderobjekts die Qualität der Übereinstimmung zwischen den Spender- und Empfängerobjekten verbessert werden kann. Aus theoretischer Sicht spre-

chen somit einige Aspekte für und einige Aspekte gegen die Begrenzung der Spenderverwendungshäufigkeit.

Im Bereich der empirischen Forschung sind in der Literatur lediglich Studien anzutreffen, die Hot-Deck-Verfahren mit anderen Imputationsverfahren vergleichen. Diese Studien betrachten ein zufälliges Ziehen der Spenderobjekte entweder nur mit Zurücklegen (vgl. z.B. Barzi, 2004, Roth, Switzer III, 1995, Yenduri, Iyengar, 2007) oder nur ohne Zurücklegen (vgl. Kaiser, 1983). Des Weiteren spielt die Frage, ob ein Hot-Deck-Verfahren mit oder ohne Begrenzung der Spenderverwendungshäufigkeit angewendet wird, bei einigen Arbeiten eine Rolle, die sich mit der Schätzung der Merkmalsvarianz auf Basis der imputierten Daten beschäftigen (vgl. Ford, 1983, Kalton, 1983, S. 25). Hier sind die beiden Fälle, ob ein Spenderobjekt mit oder ohne Zurücklegen gezogen wird, zum Teil wichtig für die Herleitung der Schätzgleichungen beziehungsweise werden in diesem Zusammenhang berücksichtigt (vgl. Brick et al., 2004).

Auf Basis des aufgezeigten Forschungsstands wird ersichtlich, dass die Notwendigkeit und die Konsequenzen einer Begrenzung der Spenderverwendungshäufigkeit noch unzureichend untersucht worden sind. Insbesondere fehlen Erkenntnisse darüber, unter welchen Konstellationen eine derartige Begrenzung sinnvoll bzw. nicht sinnvoll erscheint.

4 Design der Simulationsstudie

Im Rahmen der Simulationsstudie soll nun untersucht werden, welche Auswirkungen eine wiederholte Verwendung des Spenderobjekts bei einer Hot-Deck-Imputation auf die vervollständigte Datenmatrix hat. Dazu werden im nächsten Abschnitt zunächst die konkreten Forschungsfragen formuliert. Im Anschluss daran erfolgt in Abschnitt 4.2 eine Darstellung der Faktoren, die einen Einfluss auf die Untersuchungsergebnisse erwarten lassen und daher in der Simulationsstudie zu variieren sind. Der Abschnitt 4.3 beschäftigt sich dann mit den Gütekriterien zur Beurteilung der Simulationsergebnisse und in Abschnitt 4.4 wird schließlich noch auf die konkrete Durchführung der Simulationsstudie eingegangen.

4.1 Forschungsfragen

Im Hinblick auf eine Untersuchung der Auswirkungen einer wiederholten Verwendung des Spenderobjekts im Rahmen einer Hot-Deck-Imputation lassen sich insgesamt die folgen-

den vier konkreten Forschungsfragen ableiten, die mit der Simulationsstudie beantwortet werden sollen:

1. Ist grundsätzlich eine Beschränkung der wiederholten Verwendung eines Spenderobjekts im Rahmen einer Hot-Deck-Imputation sinnvoll bzw. notwendig?
2. Von welchen Gegebenheiten des vorliegenden Datenmaterials hängt eine notwendige Beschränkung ab?
3. Ist eine Beschränkung des Spenderobjekts abhängig vom verwendeten Hot-Deck-Verfahren?
4. Können in den jeweils betrachteten Fällen Empfehlungen hinsichtlich der Anzahl von möglichen wiederholten Verwendungen eines Spenderobjekts gegeben werden?

Wie anhand der aufgezeigten Fragestellungen zu sehen ist, soll zunächst untersucht werden, ob die Anzahl der wiederholten Verwendung eines Spenderobjekts überhaupt Auswirkungen auf die anschließende Analyse der vervollständigten Daten hat. Falls dies gezeigt werden kann, sind anschließend die Zusammenhänge mit dem vorliegenden Daten und dem verwendeten Hot-Deck-Verfahren zu untersuchen. Darüber hinaus ist dann noch zu klären, welche Empfehlungen bezüglich der Anzahl von Wiederholungen abgeleitet werden können.

4.2 Einflussfaktoren

Auf Basis der Sichtung thematisch ähnlich gelagerter Studien (vgl. Roth, 1999, Roth, Switzer III, 1995, Strike et al., 2001) sowie ergänzender Überlegungen können eine Reihe möglicher Faktoren identifiziert werden, die im Zusammenhang mit einer notwendigen Beschränkung des Spenderobjekts im Rahmen einer Hot-Deck-Imputation gegebenenfalls eine Rolle spielen. Diese Faktoren werden folglich in der Simulationsstudie variiert, um dadurch entsprechende Einflüsse analysieren zu können. Im Einzelnen sind folgende Einflussfaktoren einschließlich der für diese Studie festgelegten Ausprägungen zu nennen:

- **Dimension der Datenmatrix:** Ausgehend von n Objekten und m Merkmalen werden vier unterschiedlich dimensionierte ($n \times m$)-Datenmatrizen mit den Dimensionen (100×9), (350×9), (500×9), und (1750×9) betrachtet. Die betrachteten Größenordnungen resultieren aufgrund der nachfolgend noch aufgeführten Vorgaben zu den betrachteten Merkmalen sowie den Imputationsklassen.

- **Skalierung der Merkmale:** Es werden gemischt skalierte Datenmatrizen mit jeweils drei nominalen, ordinalen und quantitativen Merkmalen betrachtet. Bei den quantitativen Merkmalen sollen realistische, nichtnegative und nach oben begrenzte Wertebereiche generiert werden. Als nominal skalierte Merkmale werden ausnahmslos dichotome Merkmale betrachtet, da grundsätzlich auch ein nominal polytomes Merkmal durch eine entsprechende Anzahl dichotomer Merkmale dargestellt werden kann. Bei den ordinal skalierten Merkmalen werden die beiden Fälle von 5 und 7 unterschiedlichen Ausprägungen unterstellt.
- **Anzahl der Imputationsklassen:** Die Imputationsklassen werden vorab bereits festgelegt, d.h. es werden Daten entsprechend der jeweils festgelegten Klassen generiert. Dazu wird mit 2 eine eher geringe Anzahl und mit 7 eine höhere Anzahl von Klassen festgelegt.
- **Anzahl der Objekte je Imputationsklasse:** Für die bereits vorab festgelegten Imputationsklassen sollen die Fälle einer geringen und einer eher höheren Anzahl von Objekten je Klasse unterschieden werden. Dabei wird zwischen einer Anzahl von 50 und 250 Objekten je Klasse unterschieden.
- **Klassenstruktur:** Für die Imputationsklassen sollen des Weiteren die Fälle einer schwachen sowie einer relativ starken Klassenstruktur unterschieden werden. Von einer starken Klassenstruktur wird ausgegangen, wenn sich die Klassen nur bis maximal 5 % überlappen und die Merkmale innerhalb der Klassen eine paarweise Korrelation von mindestens 0,5 aufweisen. Demgegenüber wird für den Fall einer schwachen Klassenstruktur unterstellt, dass die Klassen eine Überlappung von mindestens 30 % besitzen und die Merkmale paarweise unkorreliert sind.
- **Anteil fehlender Daten:** Hier werden die Fälle 5 %, 10 % und 20 % fehlender Werte bezogen auf die gesamte Datenmatrix unterschieden.
- **Ausfallmechanismus:** Es werden die zwei unsystematischen Ausfallmechanismen MCAR und MAR sowie der Fall NMAR unterschieden.
- **Hot-Deck-Verfahren:** In der Simulationsstudie werden drei sequentielle und drei simultane Varianten von Hot-Deck-Verfahren betrachtet. Bei den sequentiellen Verfahren werden merkmalsweise die fehlenden Werte durch entsprechende Werte eines Spenderobjekts imputiert. Demgegenüber erfolgt bei den simultanen Varianten eine Imputation aller fehlenden Werte eines Objekts durch die entsprechenden

Werte des Spenderobjekts. Folglich kommen als Spenderobjekte bei den sequentiellen Verfahren jeweils alle Objekte in Frage, die bezüglich des gerade zu imputierenden Merkmals einen vorhandenen Wert besitzen, während bei den simultanen Verfahren ein Objekt nur dann als Spenderobjekt fungieren kann, wenn es vorhandenen Daten bezüglich aller zu imputierenden Merkmale aufweist. Die drei verwendeten sequentiellen sowie simultanen Verfahren unterscheiden sich jeweils darin, dass die Auswahl des Spenderobjekts zufällig (Verfahrensbezeichnungen SeqZ und SimZ) sowie gemäß der geringsten Distanz erfolgt, wobei jeweils zwei verschiedene Distanzermittlungen durchgeführt werden. Bei der ersten Distanzbestimmung wird eine linearhomogene Aggregation auf Basis der paarweise vorhandenen Daten durchgeführt (Verfahrensbezeichnungen SeqD und SimD). Bei der zweiten Distanzbestimmung erfolgt eine linearhomogene Aggregation nach einer zuvor erfolgten Imputation mit Hilfe des Lageparameters (Verfahrensbezeichnungen SeqDL und SimDL). Zur Bestimmung der merkmalsweisen Distanzen werden bei den quantitativen Merkmalen die City-Block-Metrik und bei ordinalen Merkmalen die Rangdifferenz herangezogen. Bei den nominalen Merkmalen wird zwischen Übereinstimmung und Nicht-Übereinstimmung der Ausprägungen unterschieden.

Neben diesen eben genannten Einflussfaktoren muss für die Studie noch festgelegt werden, in welchem Rahmen die wiederholte Verwendung des Spenderobjekts variiert wird. Neben den beiden Extremfällen, dass ein Objekt maximal einmal als Spender herangezogen wird sowie beliebig oft herangezogen werden kann, sollen noch zwei weitere Zwischenstufen betrachtet werden, so dass insgesamt die folgenden vier Fälle unterschieden werden:

- Ein Spenderobjekt wird nur einmal zur Imputation zugelassen.
- Ein Spenderobjekt wird maximal in 25 % der Fälle, in denen es eingesetzt werden könnte, zur Imputation zugelassen.
- Ein Spenderobjekt wird maximal in 50 % der Fälle, in denen es eingesetzt werden könnte, zur Imputation zugelassen.
- Ein Spenderobjekt kann beliebig oft zur Imputation verwendet werden.

4.3 Gütekriterien

Zur Beurteilung der Güte der durchgeführten Imputationen werden grundsätzlich die relativen Abweichungen von Verteilungsparametern der imputierten Datenmatrizen zu den entsprechenden Parametern der wahren Datenmatrizen berechnet. Abhängig von der Skalierung der Merkmale werden dabei folgende Verteilungsparameter verwendet:

- **Dichotomes Merkmal:** Ausprägungshäufigkeit
- **Ordinales Merkmal:** Median, Quartilsabstand
- **Kardinales Merkmal:** Mittelwert, Varianz

Zur Untersuchung der Unterschiede der vier Fälle einer Begrenzung der Spenderobjekte im Hinblick auf die Güte der resultierten Imputationen werden dann die Abweichungen der jeweiligen Verteilungsparameter analysiert. Für jeden der vier Fälle wird für jeden betrachteten Verteilungsparameter dazu zunächst die relative Abweichung Δp zwischen dem jeweils wahren Verteilungsparameter p_W und dem entsprechenden, auf Basis der imputierten Daten ermittelten Verteilungsparameter p_I gemäß

$$\Delta p = \frac{p_I - p_W}{p_W} \tag{1}$$

berechnet. Ein negativer Wert für Δp deutet dabei eine Unterschätzung, ein positiver Wert eine entsprechende Überschätzung des jeweiligen wahren Parameters an.

Um die Auswirkungen der unterschiedlichen Fälle einer Begrenzung der Spenderobjekte auf die Güte der Imputationen zu untersuchen, kann nun ein Vergleich der jeweils resultierenden Werte Δp durchgeführt werden. Aufgrund der in der Simulationsstudie zu erwartenden großen Datenmengen werden jedoch Signifikanztests nicht zielführend sein, so dass alternativ dazu Cohen's Effektstärke herangezogen wird (vgl. Cohen, 1992, S. 157, Bortz, Döring, 2009, S. 606). Diese ergibt sich gemäß

$$d = \frac{\left|\Delta\bar{p}_1\right| - \left|\Delta\bar{p}_2\right|}{\sqrt{\frac{s_1^2 + s_2^2}{2}}}, \tag{2}$$

wobei $\Delta\bar{p}_1$ und $\Delta\bar{p}_2$ die Mittelwerte aller nach (1) berechneten relativen Abweichungen für die zwei zu vergleichenden Fälle einer Begrenzung des Spenderobjekts bezeichnen und s_1^2 bzw. s_2^2 die zugehörigen Varianzen darstellen. Aufgrund der gleichen Stichprobenumfänge ist es nicht notwendig, eine gepoolte Standardabweichung heranzuziehen. Durch die

Verwendung der absoluten Werte für $\Delta\bar{p}_1$ und $\Delta\bar{p}_2$ in (2) kann auch das Vorzeichen von *d* interpretiert werden. Ein positives Vorzeichen bedeutet, dass der zweite Fall einer Begrenzung des Spenderobjekts besser ist als der erste Fall, während dies bei einem negativen Vorzeichen gerade umgekehrt ist. Was die Höhe von *d* betrifft, ab der man von einem bedeutsamen Effekt spricht, gibt es in der Literatur keine einheitliche Empfehlung. Gemäß Cohen (1992, S. 157) sind absolute Effektstärken von 0,2 bedeutsam, während beispielsweise Fröhlich und Pieter (2009, S. 141) bereits Werte ab 0,1 in diese Kategorie einstufen.

4.4 Durchführung der Studie

Für jede Kombination der festgelegten Faktoren „Anzahl der Imputationsklassen“, „Objektanzahl je Imputationsklasse“, „Klassenstruktur“ und „Anzahl der Ausprägungen der ordinalen Merkmale“ werden jeweils 100 vollständige Datenmatrizen erzeugt und die entsprechenden wahren Verteilungsparameter berechnet. Anschließend werden in jeder dieser 1.600 Datenmatrizen fehlende Daten generiert, wobei jede Kombination der Faktoren „Anteil fehlender Daten“ und „Ausfallmechanismus“ jeweils zehnmal herangezogen wird. Auf Basis dieser 144.000 unvollständigen Datenmatrizen erfolgt dann für jede der vier festgelegten Begrenzungen der Spenderobjekte eine Imputation mit Hilfe der sechs festgelegten Hot-Deck-Varianten, so dass letztendlich im Rahmen der Simulationsstudie 3.456.000 imputierte Datenmatrizen erzeugt werden. Für diese Datenmatrizen werden abschließend noch die entsprechenden Verteilungsparameter berechnet, um einen Vergleich mit den wahren Werten durchführen zu können.

Bei der Generierung der fehlenden Daten werden die betrachteten Ausfallmechanismen im Rahmen der Studie wie folgt simuliert: Bei MCAR werden die zu löschenden Daten durch zufälliges Ziehen ohne Zurücklegen festgelegt. Demgegenüber wird der Fall MAR dadurch erzeugt, dass abhängig von der Ausprägung des ersten dichotomen Merkmals, bei dem dann keine fehlenden Daten generiert werden, der Anteil fehlender Daten bei den anderen Merkmalen um 10 % erhöht bzw. reduziert wird. Um schließlich den Fall NMAR zu simulieren, wird die bei MAR gewählte Vorgehensweise in der Form modifiziert, dass jetzt auch beim ersten dichotomen Merkmal fehlende Daten entsprechend erzeugt werden.

Des Weiteren müssen bei der Generierung der fehlenden Daten noch Einschränkungen vorgenommen werden, um Komplikationen bei der anschließenden Imputation vorzubeugen. Zunächst muss ausgeschlossen werden, dass innerhalb einer Imputationsklasse weniger als 50 % der Objekte fehlende Daten aufweisen. Diese Einschränkung ist notwendig,

um auch den Fall problemlos abbilden zu können, dass ein Spenderobjekt nur einmal zur Imputation verwendet wird. Des Weiteren wird der Fall ausgeschlossen, dass bei einem Objekt gleichzeitig die Werte bei allen betrachteten Merkmalen fehlen.

5 Ergebnisse der Simulationsstudie

Basierend auf den Resultaten der durchgeführten Simulation sollen nun die im vorherigen Kapitel aufgezeigten Forschungsfragen beantwortet werden. Im ersten Abschnitt wird dazu zunächst untersucht, ob eine Beschränkung der Spenderverwendungshäufigkeit grundsätzlich sinnvoll ist und welche Auswirkungen diese hat. Der nachfolgenden Abschnitt 5.2 beschäftigt sich dann mit einer Analyse der Einflussfaktoren, die für oder gegen eine Beschränkung der Häufigkeit einer Verwendung von Spenderobjekten sprechen. Abschließend wird in Abschnitt 5.3 noch untersucht, inwieweit Empfehlungen hinsichtlich der Anzahl von wiederholten Verwendungen der Spenderobjekte abgeleitet werden können.

5.1 Auswirkungen einer beschränkten Verwendung von Spenderobjekten

Um die Einflussfaktoren hinsichtlich der Häufigkeit einer Verwendung der Spenderobjekte grundsätzlich zu analysieren, werden zunächst die Effektstärken nach Cohen zwischen dem Fall, dass ein Objekt maximal einmal als Spenderobjekt herangezogen wird, und dem Fall einer unbeschränkten Verwendung betrachtet. Für diese Effektstärken bezüglich der einzelnen, in Abschnitt 4.3 festgelegten Verteilungsparameter werden dann jeweils der Median sowie einige Streuungsparameter berechnet, die der Tabelle 1 entnommen werden können.

	Kardinale Merkmale		Ordinale Merkmale		Nominale Merkmale
	Mittelwert	Varianz	Median	Quartilsabstand	Ausprägungshäufigkeit
Median	-0,001	-0,009	-0,002	-0,008	-0,007
Spannweite	0,104	2,468	0,171	2,231	3,333
Abstand 90%/10%-Quantil	0,031	0,336	0,037	0,262	0,251
Quartilsabstand	0,013	0,068	0,016	0,050	0,045
Standardabweichung	0,013	0,280	0,017	0,249	0,325

Tabelle 1: Median und Variabilität der Effektstärken

Bei Betrachtung dieser Werte fällt zunächst auf, dass keine durchweg positiven oder negativen Effektstärken vorliegen, sondern die Effektstärken um den Nullpunkt in beide Richtungen streuen. Des Weiteren ist ersichtlich, dass alle Streuungsparameter bei den Vertei-

lungsparametern Mittelwert und Median relativ klein sind, so dass bezüglich dieser Verteilungsparameter nicht mit bedeutenden Effekten zu rechnen ist. Im Vergleich dazu sind bei den Verteilungsparametern Varianz, Quartilsabstand und Ausprägungshäufigkeit sehr hohe Spannweiten für die Effektstärken zu beobachten. Auch der Abstand zwischen 90%- und 10%-Quantil sowie die Standardabweichung sind bei diesen Verteilungsparametern relativ hoch. Allerdings wird bei Betrachtung des Quartilsabstands in Kombination mit dem Median auch deutlich, dass ein Großteil der Effekte als trivial eingestuft werden muss. Daraus ist insgesamt ersichtlich, dass die Verwendungshäufigkeit der Spenderobjekte grundsätzlich zwar bedeutsame Einflüsse auf die Güte der Imputation hat, dieser Effekt aber nicht durchgängig auftritt. Somit wird es von Interesse sein zu untersuchen, in welchen Situationen bedeutsame Effekte vorliegen, worauf im nächsten Abschnitt noch eingegangen wird.

Der auf Basis kombinatorischer Überlegungen resultierende Stand der Forschung, dass im Fall einer maximal einmaligen Verwendung eines Spenderobjekts die Varianz der geschätzten Verteilungsparameter der vervollständigten Daten reduziert wird, soll an dieser Stelle auf Basis der Simulationsergebnisse noch empirisch untersucht werden. In der nachfolgenden Tabelle 2 sind dazu die prozentualen Häufigkeiten angegeben, in wie vielen Fällen einer entsprechenden Parameterschätzung innerhalb einer Faktorstufenkombination eine der vier durchgeführten Spenderhäufigkeitsbegrenzungen zur kleinsten Varianz der Schätzwerte führt.

Betrachteter Parameter		Spenderhäufigkeitsbegrenzung			
		einmal	25%	50%	unbegrenzt
Kardinales Merkmal	Mittelwert	68,52%	15,47%	7,95%	8,06%
	Varianz	67,25%	15,74%	8,56%	8,45%
Ordinales Merkmal	Median	74,54%	11,38%	7,62%	6,46%
	Quartilsabstand	85,88%	5,71%	4,96%	3,45%
Dichotomes Merkmal	Ausprägungs-häufigkeit	78,36%	8,41%	6,96%	6,27%

Tabelle 2: Häufigkeitsverteilung der minimalen Varianz der Schätzwerte

Es ist deutlich zu erkennen, dass in den meisten Fällen eine maximal einmalige Verwendung eines Spenderobjekts zur kleinsten Variabilität der geschätzten Parameter führt und es damit zu einer Verbesserung der Schätzgenauigkeit kommt. Dieser Sachverhalt kommt bei den ordinalen und dichotomen Merkmalen noch stärker zum Ausdruck als bei den kardinalen Merkmalen. Allerding ist auch festzuhalten, dass eine häufigere bzw. nicht eingeschränkte Verwendung der Spenderobjekte in einer Reihe von Faktorstufenkombinationen durchaus eine minimale Variabilität des geschätzten Parameters zur Folge hat.

5.2 Analyse der Einflüsse auf die Spenderverwendungshäufigkeit

Um die Einflussfaktoren hinsichtlich der Häufigkeit einer Verwendung der Spenderobjekte grundsätzlich zu analysieren, werden wiederum die Effektstärken nach Cohen zwischen den Fällen einer einmaligen und unbegrenzten Verwendung der Spenderobjekte betrachtet. Negative Werte sprechen somit für eine Beschränkung, während positive Werte andeuten, dass eine unbeschränkte Verwendungshäufigkeit von Vorteil ist. Bei der nachfolgenden Darstellung der Ergebnisse werden zunächst die Haupteffekte und darauf aufbauend gegebenenfalls vorliegende Wechselwirkungen der Einflussfaktoren untersucht.

5.2.1 Analyse der Haupteffekte

In der Tabelle 3 sind die Effektstärken für die betrachteten Verteilungsparameter in Abhängigkeit der in dieser Studie betrachteten Einflussfaktoren zusammengefasst. Effektstärken, die im Betrag einen Wert ab 0,1 aufweisen, sind fett hervorgehoben.

		Kardinale Merkmale		Ordinale Merkmale		Nominale Merkmale
		Mittelwert	Varianz	Median	Quartilsabstand	Ausprägungshäufigkeit
Dimension der Datenmatrix	(100x9)	0,000	-0,082	-0,001	-0,030	-0,034
	(350x9)	0,000	**-0,177**	-0,005	**-0,152**	-0,022
	(500x9)	0,000	-0,064	-0,004	-0,030	**-0,130**
	(1750x9)	0,001	**-0,146**	-0,006	-0,065	**-0,162**
Anzahl der Imputationsklassen	2	0,000	-0,068	-0,001	-0,029	-0,072
	7	0,000	**-0,147**	-0,003	**-0,115**	-0,090
Objekte je Imputationsklasse	50	0,000	**-0,112**	-0,001	-0,073	-0,028
	250	0,000	-0,090	-0,005	-0,041	**-0,141**
Klassenstruktur	stark	0,000	-0,092	-0,001	-0,072	-0,072
	schwach	0,000	-0,094	-0,001	-0,045	-0,080
Prozentsatz fehlender Daten	5%	0,000	-0,025	0,000	-0,013	-0,011
	10%	0,000	-0,071	0,000	-0,037	-0,051
	20%	0,000	**-0,148**	0,000	**-0,100**	**-0,129**
Ausfallmechanismus	MCAR	0,001	-0,088	-0,001	-0,053	-0,065
	MAR	0,000	**-0,100**	0,000	-0,066	-0,086
	NMAR	0,001	-0,091	0,000	-0,058	-0,077
Hot-Deck-Verfahren	SimD	-0,001	**0,153**	-0,002	0,025	0,075
	SimDL	-0,004	**-0,339**	0,005	**-0,214**	**-0,338**
	SeqD	0,001	-0,007	-0,003	0,000	-0,005
	SeqDL	0,000	-0,088	0,010	**-0,133**	-0,041
	SimZ	0,000	-0,001	-0,001	-0,004	0,000
	SeqZ	0,000	-0,001	0,000	-0,001	-0,003

Tabelle 3: Effektstärken in Abhängigkeit der Einflussfaktoren

Bei einer Betrachtung der Ergebnisse fällt zunächst auf, dass bei den Lageparametern für kardinale und ordinale Merkmale auch unabhängig von den betrachteten Einflussgrößen nur triviale Effekte vorliegen, d.h. die Tatsache, ob ein Spenderobjekt nur einmal oder unbegrenzt zur Imputation zugelassen wird, hat für keine Ausprägung der Faktoren einen bedeutsamen Einfluss auf die Schätzgenauigkeit dieser Verteilungsparameter. Dieses Ergebnis deckt sich auch mit den Erkenntnissen aus dem vorherigen Abschnitt.

Demgegenüber können bei den Streuungsparametern sowie der Ausprägungshäufigkeit bei einigen Faktoren bedeutsame Effekte festgestellt werden. Dabei fällt auf, dass bei einem hohen Prozentsatz fehlender Daten sowie der Hot-Deck-Variante SimDL die Beschränkung der Verwendung des Spenderobjekts durchgängig zu besseren Schätzergebnissen führt. Auch eine höhere Anzahl von Imputationsklassen spricht tendenziell für eine Beschränkung.

Während die Effekte der Dimension der Datenmatrix sowie der Objektanzahl je Imputationsklasse nicht eindeutig sind, spielt es hinsichtlich der vorliegenden Klassenstruktur sowie bei Verwendung der zufälligen Hot-Deck-Varianten und der Variante SeqD keine Rolle, ob eine Beschränkung der Spenderverwendungshäufigkeit erfolgt oder nicht. Auffällig ist noch die Tatsache, dass die Hot-Deck-Variante SimD teilweise positive Effektgrößen aufweist, die für eine unbegrenzte Verwendungsmöglichkeit der Spenderobjekte sprechen. Hier könnte die Analyse von Wechselwirkungseffekten zu weiteren interessanten Erkenntnissen führen.

5.2.2 Analyse von Wechselwirkungen

Auf Basis der aus der Analyse der Haupteffekte gewonnenen Erkenntnisse sollen nun zunächst die Effektstärken für die Verteilungsparameter in Abhängigkeit aller Kombinationen zwischen den Hot-Deck-Varianten SimD, SimDL und SeqDL sowie den restlichen Einflussfaktoren untersucht werden. Die entsprechenden Werte sind dazu in der Tabelle 4 dargestellt, wobei Effektstärken ab 0,1 wiederum hervorgehoben sind.

Wie bei der Analyse der Haupteffekte zeigt sich auch jetzt, dass beim Verfahren SimD eine unbegrenzte Spenderverwendungshäufigkeit von Vorteil ist. Über alle Kombinationen mit den anderen Faktoren ergeben sich mit einer Ausnahmen positive Werte, wenngleich nur bei der Varianz und der Ausprägungshäufigkeit bedeutsame Effekte vorliegen. Des Weiteren ist zu erkennen, dass bei den beiden Verfahren SimDL und SeqDL ausnahmslos

negative Werte auftreten, die darüber hinaus größtenteils auf bedeutsame Effekte hinweisen und damit für eine einmalige Verwendung der Spenderobjekte sprechen.

		SimD			SimDL			SeqDL		
		V	Q	A	V	Q	A	V	Q	A
Dimension der Datenmatrix	(100x9)	**0,140**	0,053	0,058	**-0,337**	**-0,192**	**-0,216**	-0,089	**-0,139**	-0,026
	(350x9)	**0,235**	0,058	0,055	**-0,473**	**-0,333**	**-0,278**	**-0,120**	**-0,207**	-0,018
	(500x9)	**0,120**	0,040	**0,111**	**-0,283**	**-0,116**	**-0,492**	-0,077	-0,064	-0,073
	(1750x9)	**0,215**	0,045	**0,108**	**-0,420**	**-0,257**	**-0,554**	**-0,109**	**-0,132**	-0,064
Klassenanzahl	2	0,097	0,025	0,081	**-0,247**	**-0,101**	**-0,300**	-0,066	-0,082	-0,049
	7	**0,287**	0,033	0,075	**-0,521**	**-0,382**	**-0,424**	**-0,130**	**-0,217**	-0,031
Objekte je Klasse	50	**0,182**	0,082	0,034	**-0,426**	**-0,284**	**-0,132**	**-0,111**	**-0,196**	-0,004
	250	**0,143**	0,056	**0,140**	**-0,319**	**-0,131**	**-0,684**	-0,088	-0,047	-0,098
Klassenstruktur	stark	**0,153**	0,048	0,078	**-0,339**	**-0,156**	**-0,362**	-0,091	**-0,135**	-0,042
	schwach	**0,144**	0,006	0,071	**-0,338**	**-0,269**	**-0,313**	-0,085	**-0,132**	-0,040
Prozentsatz fehlender Daten	5%	0,065	-0,012	0,031	-0,084	-0,057	-0,045	-0,013	-0,028	-0,004
	10%	**0,148**	0,006	0,077	**-0,262**	**-0,162**	**-0,213**	-0,039	-0,073	-0,010
	20%	**0,203**	0,061	**0,101**	**-0,558**	**-0,345**	**-0,600**	**-0,168**	**-0,233**	-0,085
Ausfallmechanismus	MAR	**0,151**	0,025	0,079	**-0,355**	**-0,226**	**-0,372**	**-0,107**	**-0,152**	-0,058
	MCAR	**0,153**	0,023	0,067	**-0,326**	**-0,204**	**-0,296**	-0,075	**-0,119**	-0,025
	NMAR	**0,154**	0,029	0,077	**-0,334**	**-0,213**	**-0,344**	-0,081	**-0,125**	-0,038

Tabelle 4: Wechselwirkungen zwischen Imputationsmethode und restlichen Faktoren (Legende: V = Varianz, Q = Quartilsabstand, A = Ausprägungshäufigkeit)

Für alle drei betrachteten Hot-Deck-Varianten ist ersichtlich, dass bei einer hohen Zahl von Imputationsklassen sowie einem höheren Prozentsatz fehlender Daten bedeutsame Effekte vorliegen. Bezüglich der Anzahl der Objekte je Imputationsklasse zeigt sich demgegenüber kein einheitlicher Effekt, da je nach Hot-Deck-Variante und Skalierung der Merkmale eine geringere wie auch eine höhere Anzahl an Objekten je Klasse einen bedeutenden Effekt hervorruft. Bei den restlichen Einflussfaktoren lassen sich unabhängig von den jeweiligen Ausprägungen bedeutsame und triviale Effekte gleichermaßen feststellen, die damit nur auf das Imputationsverfahren bzw. die Skalierung der Merkmale zurückzuführen sind.

Neben den nach den Hot-Deck-Varianten differenziert betrachteten Einflüssen liegen auch einige Wechselwirkungen höherer Ordnung vor, die zu auffällig großen absoluten Werten für die Effektstärke führen. Beispielsweise treten im Fall 20 % fehlender Werte sowie einer hohen Anzahl von Imputationsklassen und einer geringen Objektanzahl in diesen Klassen bei dem Verfahren SimDL Effektstärken bis zu -1,7 für die Varianz und bis zu -1,9 beim Quartilsabstand auf. Effektstärken bis zu einem Wert von -3 ergeben sich für die Ausprägungshäufigkeit im Fall einer hohen Klassenanzahl mit vielen Objekten je Klasse und ei-

nem hohen Anteil fehlender Daten. Betrachtet man demgegenüber das Verfahren SimD, so sind die stärksten Effekte im Fall einer hohen Klassenanzahl mit wenigen Objekten je Klasse und einem hohen Anteil fehlender Daten deutlich kleiner mit Werten von maximal 0,6 bei der Varianz und 0,34 beim Quartilsabstand. Auffällig ist dennoch, dass insbesondere die Kombination aus Hot-Deck-Variante, Anzahl der Imputationsklassen, Objekte je Klasse und Anteil fehlender Daten sehr bedeutsame Effektstärken zur Folge haben, die für eine maximal einmalige wie auch unbeschränkte Verwendungshäufigkeit des Spenderobjekts sprechen. Darüber hinaus bestätigt die Analyse der Wechselwirkungen höherer Ordnung durchgängig die für einzelne Hot-Deck-Verfahren bereits festgestellte, jeweils vorteilhafte Form einer Begrenzung der Spenderverwendungshäufigkeit.

5.3 Analyse der Häufigkeit einer Verwendung der Spenderobjekte

Bislang wurde im Wesentlichen untersucht, ob und unter welchen Bedingungen eine einmalige oder eine unbeschränkte Verwendung des Spenderobjekts zu besseren Ergebnissen führt. Jetzt sollen auch Spenderhäufigkeitsbegrenzungen zwischen diesen beiden Extremfällen in die Betrachtung aufgenommen werden. Dazu wird ermittelt, wie häufig einer der in dieser Studie betrachteten vier Fälle einer Beschränkung der Spenderverwendungshäufigkeit die beste Parameterschätzung liefert. Die entsprechenden prozentualen Häufigkeiten sind in der nachfolgenden Tabelle 5 dargestellt.

Betrachteter Parameter		Spenderhäufigkeitsbegrenzung			
		einmal	25%	50%	unbegrenzt
Kardinales Merkmal	Mittelwert	42,71%	20,22%	18,48%	18,60%
	Varianz	54,05%	17,79%	13,04%	15,12%
Ordinales Merkmal	Median	46,41%	21,53%	14,47%	17,59%
	Quartilsabstand	56,83%	16,24%	12,94%	13,99%
Dichotomes Merkmal	Ausprägungs-häufigkeit	49,42%	18,94%	15,07%	16,57%

Tabelle 5: Häufigkeitsverteilung der geringsten Abweichung vom wahren Verteilungsparameter

Es zeigt sich, dass in den meisten Fällen eine maximal einmalige Verwendung eines Spenderobjekts zur besten Parameterschätzung führt. Dieser Sachverhalt kommt bei den Variabilitätsmaßen stärker als bei den Lageparametern zum Ausdruck. Dabei ist durchgängig zu erkennen, dass die Häufigkeiten über die vier betrachteten Fälle einer Spenderhäufigkeitsbegrenzung zunächst abnehmen und danach wieder zunehmen, so dass nochmals deutlich wird, dass situationsbedingt unterschiedliche Beschränkungen jeweils von Vorteil sind.

Auf Basis dieser Ergebnisse kann festgehalten werden, dass grundsätzlich ein Optimierungspotenzial bezüglich der konkreten Anzahl, wie häufig ein Spenderobjekt maximal herangezogen werden soll, erkennbar ist. An dieser Stelle macht es jedoch keinen Sinn, die zugrundeliegenden Zusammenhänge näher zu analysieren und konkrete Empfehlungen diesbezüglich abzuleiten, da in dieser Studie nur vier Fälle einer Spenderhäufigkeitsbegrenzung unterschieden worden sind. Die hier gewonnenen Erkenntnisse sprechen jedoch eindeutig dafür, entsprechende zukünftige Forschungsbemühungen auf diesen Bereich zu fokussieren.

6 Zusammenfassung und Ausblick

Die im Rahmen dieser Arbeit durchgeführte Simulationsstudie zeigt, dass es deutliche Unterschiede zwischen Hot-Deck-Imputationen gibt, bei denen die Spenderverwendungshäufigkeit variiert wird. Eine Beschränkung der wiederholten Verwendung eines Spenderobjekts ist nicht grundsätzlich von Vorteil, da situationsbedingt auch eine unbeschränkte Spenderverwendung zu besseren Ergebnissen führen kann.

Einige Gegebenheiten des vorliegenden Datenmaterials sprechen im Hinblick auf dadurch verbesserten Parameterschätzungen für eine Einschränkung der Spenderverwendungshäufigkeit. Falls die Anzahl der Imputationsklassen gering ist, ergibt sich ein Vorteil hinsichtlich der Schätzgenauigkeit der Streuungsparameter bei kardinalen und ordinalen Merkmalen. Bei eher wenigen Objekten je Imputationsklasse profitiert die Varianz der kardinalen Merkmale von einer Begrenzung, während bei den dichotomen Merkmalen viele Objekte für eine Begrenzung sprechen. Darüber hinaus ist dies grundsätzlich bei einer hohen Anzahl von fehlenden Daten der Fall. Insgesamt kann auch festgehalten werden, dass die Schätzung der Lageparameter der kardinalen und ordinalen Merkmale nicht nennenswert durch eine Begrenzung der Spenderverwendungshäufigkeit beeinflusst wird.

Neben den Gegebenheiten des Datenmaterials stellt vor allem die verwendete Hot-Deck-Variante einen bedeutenden Einflussfaktor dar. Je nach Verfahrensvariante bringt eine Begrenzung der Spenderverwendungshäufigkeit Vorteile oder Nachteile mit sich bzw. ist gegebenenfalls auch ohne nennenswerten Einfluss auf die Parameterschätzungen. Bei den beiden zufälligen Imputationsverfahren und der Variante SeqD zeigt eine Begrenzung nie bedeutsame Effekte. Demgegenüber ist im Fall der Verfahrensvarianten SeqDL und

SimDL eine Begrenzung sinnvoll, während bei SimD eine unbegrenzte Verwendung der Spenderobjekte zu empfehlen ist.

Auch wenn in den meisten Fällen eine maximal einmalige Verwendung eines Spenderobjekts zur besten Parameterschätzung führt, sind situationsbedingt auch weniger restriktive bzw. keine Beschränkungen häufig von Vorteil. Die Bestimmung einer auf den konkreten Fall angepassten Grenze der Spenderverwendungshäufigkeit erscheint somit sinnvoll, so dass detaillierte Untersuchungen der diesbezüglich zugrundeliegenden Zusammenhänge einen sehr interessanten Ansatz für zukünftige Forschungsarbeiten darstellen könnten.

Literaturverzeichnis

Allison, P.D. (2001): Missing Data, Sage University Papers Series on Quantitative Applications in the Social Sciences, Thousand Oaks

Andridge, R.R., Little, R.J.A. (2010): A Review of Hot Deck Imputation for Survey Nonresponse, International Statistical Review, 78, 1, S. 40-64

Bankhofer, U. (1995): Unvollständige Daten- und Distanzmatrizen in der Multivariaten Datenanalyse, Eul, Bergisch Gladbach

Barzi, F., Woodward, M. (2004): Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies, American Journal of Epidemiology, 160, S. 34-45

Borz, J., Döring, N. (2009): Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler, Springer, Berlin

Brick, J.M., Kalton, G. (1996): Handling Missing Data in Survey Research, Statistical Methods in Medical Research, 5, S. 215-238

Brick, J.M., Kalton, G., Kim, J.K. (2004): Variance Estimation with Hot Deck Imputation Using a Model, Survey Methodology, 30, S. 57-66

Cohen, J. (1992): A Power Primer, Quantitative Methods in Psychology, 112, S. 155-159

Fröhlich, M., Pieter, A. (2009): Cohen's Effektstärken als Mass der Bewertung von praktischer Relevanz – Implikationen für die Praxis, Schweizerische Zeitschrift für Sportmedizin und Sporttraumatologie, 57, 4, S. 139-142

Ford, B. (1983): An Overview of Hot-Deck Procedures, Madow, W., Nisselson, H., Olkin, I. (Hrsg.), Incomplete Data in Sample Surveys, 2, Theory and Bibliographies, Academic Press, S. 185-207

Kaiser, J. (1983): The Effectiveness of Hot-Deck Procedures in Small Samples, Proceedings of the Section on Survey Research Methods, American Statistical Association, S. 523-528

Kalton, G. (1983): Compensating for Missing Survey Data, Ann Arbor: Institute for Social Research, University of Michigan

Kalton, G., Kasprzyk, D. (1982): Imputing for Missing Survey Responses, Proceedings of the Section on Survey Research Methods, American Statistical Association, S. 22-31

Kalton, G., Kasprzyk, D. (1986): The Treatment of Missing Survey Data, Survey Methodology, 12, S. 1-16

Kalton, G., Kish, L. (1981): Two Efficient Random Imputation Procedures, Proceedings of the Survey Research Methods Section 1981, S. 146-151

Kim, J.O., Curry, J. (1977): The Treatment of Missing Data in Multivariate Analysis, Sociological Methods and Research, 6, S. 215-240

Little, R.J., Rubin, D.B. (1987): Statistical Analysis with Missing Data, New York, Wiley

Marker, D.A., Judkins, D.R., Winglee, M. (2002): Large-scale Imputation for Complex surveys, Groves, R.M., Dillman, D.A., Eltinge, J.L., Little, R.J.A. (Hrsg.), Survey Nonresponse, Chapter 22, New York, Wiley

Roth, P.L. (1999): Missing Data in Multiple Item Scales: A Monte Carlo Analysis of Missing Data Techniques, Organizational Research Methods, 2, S. 211-232

Roth, P.L., Switzer III, F.S. (1995): A Monte Carlo Analysis of Missing Data Techniques in a HRM Setting, Journal of Management, 21, S. 1003-1023

Sande, I. (1983): Hot-Deck Imputation Procedures, Madow, W., Nisselson, H., Olkin, I. (Hrsg.), Incomplete Data in Sample Surveys, 3, Theory and Bibliographies, Academic Press, S. 339-349

Schnell, R. (1986): Missing-Data Problem in der empirischen Sozialforschung, Dissertation, Bochum

Strike, K., Emam, K.E., Madhavji, N. (2001): Software Cost Estimation with Incomplete Data, IEEE Transactions on Software Engineering, 27, S. 890-908

Yenduri, S., Iyengar, S.S. (2007): Performance Evaluation of Imputation Methods for Incomplete Datasets, International Journal of Software Engineering and Knowledge Engineering, 17, S. 127-152